

# **BAB I**

## **PENDAHULUAN**

### **1.1 Latar Belakang**

Baseball adalah salah satu olahraga yang menggunakan statistik. Mulai dari statistik tradisional seperti Hit, Run, Out, dan Error hingga statistik lanjutan seperti Wins Above Replacement (WAR) dan Annual Average Value (AAV) yang dapat digunakan untuk mengukur kinerja suatu pemain dan berapa harga kontrak yang pantas untuk dibayarkan kepada pemain tersebut. Hal ini dimulai oleh Henry Chadwick pada tahun 1858, di mana ia menemukan box scores yaitu tabel yang berisi catatan permainan dan skor serta statistik para pemain[1]. Box scores membuahakan statistik untuk perkembangan dalam olahraga. Banyak analis dan penggemar baseball menggunakan record yang ada untuk membangun model prediktif berdasarkan kategori statistik baru dari data yang dihasilkan pada permainan tersebut.

Baseball sendiri memiliki 121 tipe data statistik yang dibagi menjadi 72 data standard dan 49 data advance. Diluar 121 tipe data tersebut ada juga 32 tipe data yang dilabeli sebagai data statcast yang merupakan automated tool yang dikembangkan untuk menganalisis pergerakan dan kemampuan atletik pemain. Statcast diperkenalkan ke 30 stadion Major League Baseball (MLB) pada tahun 2015[2].

Data yang ada dapat diperluas ke metrik kinerja tim. Misalnya, analis dapat memetakan kemampuan bertahan tim untuk membuat runner on base

mendapatkan out dari berbagai titik di lapangan, dengan memperhitungkan efisiensi dan kecepatan lemparan oleh Catcher. Data statistik pada baseball juga dapat digunakan untuk mencegah cedera dengan melacak metrik kinerja fisik sepanjang musim.

Data statistik yang melimpah ini memungkinkan pengembangan model prediktif untuk mengukur performa pemain atau tim. Model prediksi seperti Linear Regression digunakan untuk menemukan hubungan sederhana antara variabel, seperti memprediksi jumlah pukulan berdasarkan waktu bermain. Namun, dengan meningkatnya kompleksitas data yang melibatkan interaksi antar variabel, algoritma seperti XGBoost (Extreme Gradient Boosting) menjadi pilihan yang lebih efektif. XGBoost unggul dalam menangani data non-linear dan besar, serta sering digunakan dalam prediksi performa pemain untuk memperkirakan variabel seperti kecepatan bola, jumlah home run, atau peluang kemenangan tim berdasarkan performa historis[3].

Selain membantu memaksimalkan performa pemain, analisis statistik baseball juga dapat digunakan untuk mencegah cedera dengan memantau metrik fisik pemain sepanjang musim[4]. Data ini juga memiliki implikasi ekonomi yang besar. Pemilik tim olahraga bukan hanya memenangkan kejuaraan tetapi juga untuk mendapatkan keuntungan saat menjalankan kompetisi. Baseball, seperti bisnis lainnya menghasilkan sejumlah besar uang yang memungkinkan tim untuk membeli pemain sesuai kebutuhannya. Menurut Statista (2023), Major League Baseball (MLB) menghasilkan pendapatan keseluruhan sebesar 10,3 miliar dollar AS pada tahun 2023[5]. Sebagai

perbandingan, National Football League (NFL) mencatat 18,6 miliar dollar AS pada tahun 2022, dan National Basketball Association (NBA) mencapai 10,5 miliar dollar AS pada tahun 2022.

Pada liga olahraga profesional, beberapa tim mengeluarkan uang lebih banyak daripada tim lain. Hal ini disebabkan oleh umur organisasi tersebut, sponsor, dan tempat mereka bermain. Seperti halnya tim yang berada di New York dan Los Angeles memiliki lebih banyak penggemar sehingga mendatangkan banyak sponsor dibandingkan tim yang berasal dari Milwaukee dan Oakland. Akibatnya, pengeluaran tim tidak dapat seimbang dengan nilai organisasi tersebut[6]. Berikut adalah 5 nilai organisasi paling berharga dan paling berkekurangan tiap cabang olahraga pada tahun 2023 :

*Tabel 1.1 Perbandingan Nilai Organisasi pada 5 tim terkaya dan termiskin pada NFL, NBA, dan MLB*

	<b>Top 5</b>	<b>Nilai Organisasi (Dalam Miliar dollar AS)</b>	<b>Mean</b>	<b>Bottom 5</b>	<b>Nilai Organisasi (Dalam Miliar dollar AS)</b>	<b>Mean</b>
<b>NFL</b>	<b>Cowboys</b>	<b>9</b>	<b>7.2</b>	<b>Bengals</b>	<b>3.5</b>	<b>3.76</b>
	<b>Patriots</b>	<b>7</b>		<b>Lions</b>	<b>3.6</b>	
	<b>Rams</b>	<b>6.9</b>		<b>Bills</b>	<b>3.7</b>	
	<b>Giants</b>	<b>6.8</b>		<b>Cardinals</b>	<b>3.8</b>	
	<b>Bears</b>	<b>6.3</b>		<b>Jaguars</b>	<b>4</b>	
<b>NBA</b>	<b>Warriors</b>	<b>7.7</b>	<b>6.01</b>	<b>Grizzlies</b>	<b>2.4</b>	<b>2.66</b>
	<b>Knicks</b>	<b>6.6</b>		<b>Timberwolves</b>	<b>2.5</b>	
	<b>Lakers</b>	<b>6.4</b>		<b>Pelicans</b>	<b>2.55</b>	
	<b>Celtics</b>	<b>4.7</b>		<b>Pacers</b>	<b>2.9</b>	
	<b>Clippers</b>	<b>4.65</b>		<b>Magic</b>	<b>2.95</b>	
<b>MLB</b>	<b>Yankees</b>	<b>7.1</b>	<b>4.84</b>	<b>Marlins</b>	<b>1</b>	<b>1.402</b>
	<b>Dodgers</b>	<b>4.8</b>		<b>Athletics</b>	<b>1.18</b>	
	<b>Red Sox</b>	<b>4.5</b>		<b>Reds</b>	<b>1.19</b>	
	<b>Cubs</b>	<b>4.1</b>		<b>Royals</b>	<b>1.2</b>	
	<b>Giants</b>	<b>3.7</b>		<b>Rays</b>	<b>1.25</b>	

Lima tim teratas di NFL rata-rata bernilai 91% lebih tinggi dibandingkan lima tim terbawah pada liga tersebut. Sebagai perbandingan rasio di NBA

sebesar 125%. Proporsi ini nampaknya cukup masuk akal karena sulitnya mempertahankan keseimbangan yang sempurna dalam liga. Namun, kemudian muncullah MLB, dengan perbandingan rasio 245% antara lima tim terkaya dan lima tim terbawah. Padahal MLB dan NBA memiliki angka pendapatan yang serupa yaitu 10,3 miliar dollar AS dan 10,5 miliar dollar AS. Poin-poin penting seperti membandingkan penilaian valuasi tim terhadap pengeluaran untuk gaji pemain yang tidak proporsional ini terjadi sehingga tim yang memiliki pengeluaran rendah harus menggunakan statistik untuk mencari pemain yang cocok untuk gaya bermain tim dengan gaji yang cukup rendah.

*Tabel 1.2 Perbandingan Pengeluaran pada 5 tim terkaya dan termiskin pada NFL, NBA, dan MLB*

	<b>Top 5</b>	<b>Nilai Organisasi (Dalam Jutaan dollar AS)</b>	<b>Mean</b>	<b>Bottom 5</b>	<b>Nilai Organisasi (Dalam Jutaan dollar AS)</b>	<b>Mean</b>
<b>NFL</b>	<b>Panthers</b>	232.7	<b>230.6</b>	<b>Niners</b>	199.5	<b>212</b>
	<b>Eagles</b>	232.5		<b>Vikings</b>	212.7	
	<b>Broncos</b>	231.7		<b>Seahawks</b>	214.9	
	<b>Bills</b>	228.6		<b>Bengals</b>	215.8	
	<b>Bears</b>	227.5		<b>Rams</b>	217.1	
<b>NBA</b>	<b>Warriors</b>	224.9	<b>202</b>	<b>Jazz</b>	130.4	<b>138.2</b>
	<b>Clippers</b>	202		<b>Pistions</b>	135.4	
	<b>Timberwolves</b>	197.8		<b>Magic</b>	139.3	
	<b>Suns</b>	196.8		<b>Spurs</b>	142.9	
	<b>Bucks</b>	188.2		<b>Hornets</b>	143	
<b>MLB</b>	<b>Mets</b>	343.6	<b>245.4</b>	<b>Athletics</b>	62.2	<b>75.9</b>
	<b>Yankees</b>	278.6		<b>Orioles</b>	71	
	<b>Padres</b>	256		<b>Pirates</b>	75.6	
	<b>Rangers</b>	251.3		<b>Rays</b>	79.3	
	<b>Phillies</b>	245.5		<b>Guardians</b>	91.8	

Selain analisis penilaian valuasi organisasi, pada Tabel 1.2 disajikan perbandingan pengeluaran gaji antar liga. Rata-rata, lima tim teratas di NFL menghabiskan 8,77% lebih banyak uang untuk gaji dibandingkan organisasi dengan pengeluaran paling sedikit. Persentase rata-rata meningkat menjadi 46.16% untuk

NBA, namun perbedaannya tidak jauh dengan MLB. Pengeluaran gaji untuk tim terkaya di MLB lebih tinggi 361% dari tim dengan pengeluaran terendah[7]. Bahkan jika kita menggabungkan 4 tim dengan pengeluaran terendah (Athletics, Orioles, Pirates, dan Rays) masih tetap tidak akan menjadikan mereka sebagai tim dengan pengeluaran tertinggi dengan selisih terhadap Mets sebesar 55,5 juta dollar AS. Perbandingan ini bertujuan untuk melihat kesenjangan uang yang cukup besar antara tim terkaya terhadap tim yang disebut “*bottom feeders*”[7].

Tim yang memiliki nilai kekayaan dan pengeluaran yang rendah harus berjuang setiap musim melawan tim-tim dengan kekayaan yang sangat tidak masuk akal. Oleh karena itu, mereka harus tahu bagaimana mengalokasikan uang mereka dengan benar. Mereka harus fokus pada dampak statistik lanjutan untuk mengevaluasi pemain[8]. Tim dapat mengevaluasi pemain berdasarkan kinerja mereka sebelumnya dan menemukan pemain dengan potensi tinggi dan cocok dengan gaya permainan tim dengan bayaran yang lebih murah[9].

Dalam industri olahraga profesional seperti Major League Baseball (MLB), penentuan gaji pemain merupakan proses yang kompleks dan strategis. Gaji tidak hanya mencerminkan kontribusi pemain di masa lalu, tetapi juga merupakan bentuk investasi terhadap potensi mereka di masa depan. Namun, proses penilaian ini sering kali subjektif dan dipengaruhi oleh banyak faktor non-performa seperti reputasi, popularitas, atau negosiasi kontrak yang tidak efisien. Di sinilah pentingnya model prediksi gaji berbasis data: untuk menghadirkan pendekatan yang objektif, terukur, dan transparan dalam

pengambilan keputusan manajemen tim terkait kontrak dan kompensasi pemain.

Permasalahan yang ingin diselesaikan melalui penelitian ini adalah ketidaksesuaian antara gaji pemain dengan performa aktual mereka, yang berisiko menyebabkan pemborosan anggaran tim atau ketimpangan dalam struktur gaji. Selain itu, tim dengan sumber daya terbatas memerlukan pendekatan berbasis data untuk membuat keputusan gaji yang cerdas dan kompetitif. Dengan adanya model prediktif, manajemen dapat memperkirakan besaran gaji yang layak untuk pemain berdasarkan indikator statistik yang relevan, seperti batting average, home run, on-base percentage, dan faktor-faktor lainnya yang mencerminkan kontribusi riil pemain terhadap tim.

Penelitian ini mengembangkan model prediksi gaji menggunakan dua algoritma: XGBoost dan Linear Regression. Pemilihan Linear Regression didasarkan pada kesederhanaan dan interpretabilitas model, yang memungkinkan pemahaman langsung tentang kontribusi masing-masing variabel terhadap nilai gaji. Sementara itu, XGBoost dipilih karena kemampuannya dalam menangani data kompleks dan menghasilkan prediksi dengan akurasi tinggi, terutama dalam kasus dengan interaksi non-linear antar variabel. Dengan membandingkan kedua pendekatan ini, penelitian ini tidak hanya menawarkan model prediksi gaji yang presisi, tetapi juga memberikan wawasan metodologis terkait efektivitas algoritma dalam konteks analisis performa olahraga.

Pemilihan topik prediksi dalam olahraga baseball dibandingkan dengan cabang olahraga lainnya didasarkan pada karakteristik baseball yang memiliki keterikatan kuat dengan data dan statistik. Berbeda dengan olahraga lain yang cenderung lebih dipengaruhi oleh elemen subjektif atau dinamika permainan yang cepat berubah, baseball menawarkan struktur permainan yang stabil, unit aksi yang jelas, serta catatan statistik historis yang terdokumentasi secara rinci [30]. Setiap aspek dalam permainan—seperti lemparan, pukulan, lari, dan tangkapan—dapat diukur secara kuantitatif, sehingga menjadikan baseball sebagai objek yang ideal untuk analisis data dan prediksi performa [31]. Selain itu, dengan berkembangnya pendekatan *sabermetrics* dalam beberapa dekade terakhir, peluang untuk menerapkan teknik analitik dan pembelajaran mesin dalam mengevaluasi performa pemain maupun kesuksesan tim semakin terbuka lebar [32]. Oleh karena itu, baseball dipandang sebagai subjek yang relevan dan potensial untuk diteliti dalam konteks prediksi berbasis data.

Secara keseluruhan, manfaat dari pengembangan model prediksi gaji ini adalah untuk memberikan alat bantu yang dapat digunakan oleh manajer tim, analis performa, maupun agen pemain dalam proses perencanaan kontrak, negosiasi, dan pengelolaan anggaran jangka panjang. Model ini juga membuka ruang bagi adopsi pendekatan *data-driven decision making* dalam dunia olahraga, yang semakin relevan di era analitik saat ini.

## 1.2 Rumusan Masalah

Berdasarkan uraian tentang latar belakang masalah yang telah disajikan sebelumnya, masalah-masalah tersebut dapat dijabarkan menjadi beberapa poin utama sebagai berikut:

1. Bagaimana mengembangkan model prediksi gaji pemain baseball untuk musim mendatang menggunakan algoritma XGBoost dan Linear Regression?

## 1.3 Batasan Masalah

Dalam penelitian ini, terdapat beberapa batasan masalah pada perusahaan yang dianalisis. Batasan-batasan tersebut adalah sebagai berikut:

1. Penelitian ini hanya menggunakan data statistik pemain bisbol dalam rentang waktu 2010 hingga 2024.
2. Data yang dipakai untuk melakukan olah data pada modelling adalah Salary Differences, Age, Hits, Runs, RBI, Walks, Strikeout, Stolen Bases, dan OPS.
3. Tidak membahas faktor eksternal yang dapat memengaruhi nilai gaji pemain seperti peran agen pemain, daya tarik pasar, dan permintaan pasar.

## 1.4 Tujuan dan Manfaat Penelitian

### 1.4.1 Tujuan Penelitian

1. Membandingkan performa kedua algoritma dalam memprediksi gaji pemain bisbol.

2. Menghasilkan model yang dapat memprediksi gaji pemain baseball dengan menggunakan statistik menggunakan algoritma XGBoost dan Linear Regression.

#### **1.4.2 Manfaat Penelitian**

Manfaat pada penelitian ini adalah sebagai berikut:

1. Bagi tim bisbol, Memprediksi nilai gaji pemain di musim Memberikan kontribusi terhadap pengembangan metode prediktif dalam analisis data olahraga, khususnya di bidang bisbol, dengan memanfaatkan algoritma modern seperti XGBoost untuk prediksi dan analisis data dan menambah wawasan tentang penerapan Linear Regression dalam evaluasi hubungan antara nilai pemain dan kontribusi kinerja mereka terhadap tim.
2. Bagi Akademisi, Penelitian ini memberikan kontribusi dalam pengembangan referensi akademik terkait penerapan algoritma machine learning dalam menganalisis data statistik pemain baseball untuk memprediksi dampaknya terhadap kesuksesan tim. Hasil penelitian ini dapat menjadi dasar bagi studi lanjutan dalam bidang analisis performa olahraga berbasis data serta pengembangan model prediktif dalam analisa statistik olahraga.

#### **1.5 Sistematika Penulisan**

Tugas akhir ini dibagi menjadi lima bab pembahasan, dengan uraian sebagai berikut:

#### **BAB I PENDAHULUAN**

Bagian ini berisi latar belakang masalah, yaitu pentingnya penggunaan data statistik dalam analisis performa pemain dan tim baseball. Fokus penelitian pada perbandingan dua algoritma, Linear Regression dan XGBoost, sebagai model prediksi. Selain itu, bagian ini juga akan membahas ruang lingkup penelitian yang melibatkan data statistik baseball, tujuan penelitian untuk menentukan algoritma yang lebih akurat dalam memprediksi performa, serta manfaat penelitian bagi tim manajer dan analis olahraga. Metode penelitian yang digunakan, termasuk pengumpulan data, preprocessing, implementasi model, dan evaluasi performa model juga akan dijelaskan. Sistematika penulisan di akhir bab memberikan gambaran umum struktur tugas akhir ini.

## BAB II LANDASAN TEORI

Bab ini memaparkan teori-teori yang mendasari penggunaan Linear Regression dan XGBoost dalam analisis prediksi. Penulis akan mengulas literatur terkait dengan penerapan model prediktif dalam olahraga, khususnya dalam baseball, serta bagaimana kedua algoritma tersebut digunakan dalam berbagai bidang, termasuk prediksi harga saham dan performa pemain olahraga. Kerangka teori ini mencakup pemahaman dasar tentang Linear Regression, XGBoost, dan analisis statistik dalam konteks olahraga. Literatur terkait penelitian terdahulu juga dibahas untuk memperkuat fondasi pemikiran yang akan diterapkan dalam penelitian ini.

## BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan pendekatan yang digunakan dalam penelitian, mulai dari pengumpulan data statistik pertandingan dan performa pemain baseball hingga pemilihan algoritma yang tepat untuk model prediksi. Metode yang digunakan

untuk membersihkan, menormalisasi, dan memisahkan data menjadi set pelatihan dan pengujian juga dijelaskan. Implementasi algoritma Linear Regression dan XGBoost diuraikan secara rinci, termasuk cara pengaturan parameter untuk masing-masing model. Selain itu, metode evaluasi performa model, seperti RMSE, dan  $R^2$ , akan dibahas untuk mengukur keakuratan kedua algoritma.

#### BAB IV ANALISIS DAN HASIL PENELITIAN

Pada bab ini, dilakukan analisis terhadap data yang telah dikumpulkan dan diolah menggunakan kedua algoritma yang diuji. Hasil implementasi model Linear Regression dan XGBoost dibandingkan untuk melihat performa masing-masing algoritma dalam memprediksi performa pemain. Analisis meliputi evaluasi metrik performa seperti akurasi, RMSE, dan  $R^2$  untuk menilai keandalan dan efisiensi model dalam konteks data baseball. Bagian ini juga mencakup hasil perbandingan antara kedua model, serta analisis mendalam terhadap kelebihan dan kekurangan masing-masing algoritma dalam memprediksi hasil yang diinginkan.

#### BAB V SIMPULAN DAN SARAN

Bab ini merangkum kesimpulan dari hasil penelitian, di mana algoritma yang lebih efektif dalam memprediksi performa pemain dan tim baseball akan dianalisis. Penulis akan memberikan rekomendasi mengenai pemilihan algoritma yang paling sesuai berdasarkan hasil perbandingan performa kedua model. Selain itu, saran untuk penelitian lebih lanjut, termasuk penggunaan data tambahan dan eksplorasi algoritma lain yang lebih kompleks, akan diberikan untuk meningkatkan akurasi model prediksi di masa depan.